Systematic Approach to Data Quality for Thermodynamic Property Databases

Q. Dong [C, S]
Physical and Chemical Property Division, National Institute of Standards and Technology, Boulder, CO, U.S.A.

X.J. Yan
Thermodynamics Research Center (TRC), Physical and Chemical Property Division, National Institute of Standards and Technology, Boulder, CO, U.S.A.

R.D. Chirico
Thermodynamics Research Center (TRC) Physical and Chemical Properties Division, National Institute of Standards and Technology, Boulder, CO, U.S.A.

V.V. Diky
Thermodynamics Research Center, National Institute of Standards and Technology, Boulder, CO, U.S.A.

M. Frenkel
Thermodynamics Research Center (TRC)
Physical and Chemical Properties Division, National Institute of Standards and Technology, Boulder, CO, U.S.A.

The modern world is an information-driven world, and we are surrounded by and dependent on data. Nevertheless, in the real world data are rife with uncertainty; databases are highly susceptible to a variety of errors, classified as anomalous, incomplete, inconsistent or erroneous data, due to the human-involving nature of the system in which data are generated. It was reported that error rates of 1-5% are typical, with an estimated immediate cost of about 10% of revenue in industry, and a similar analysis of error rate for scientific databases was made as well. Data quality problems are exacerbated in large-scale databases where data are collected from various sources over a long period of time without implementing effective data quality assurance measures. How can a large set of data be cleansed? How can data quality measures be an effective part of database operation and management system? What are the important concepts and techniques involved? This presentation addresses these issues, based on an effort to implement a systematic approach of data quality assurance in the NIST/TRC Source data system, which is an extensive repository system of experimental thermophysical and thermochemical property data that have been reported in the world's scientific literature. Data quality assurance is not just data cleansing and correction; it is also identifying the sources of poor quality and fixing the process to prevent recurrence, which require integration of techniques and methods in database management, statistical data analysis, and scientific principles. In addition to conventional data integrity constrains employed in RDBMS, scientific data integrity is utilized as criteria to detect anomalous values of numerical data, in which basic thermodynamic principles, structures of substances, and predictive models provide a variety of data consistency checks and cluster analysis. The anomaly detection helps to comprehend the overall quality of a particular set of properties or of the database, and to determine some specific strategy of combined computer and human inspection. A systematic approach of data quality assurance is developed to span the life cycle of database, including error prevention, data analysis, database integrity enforcement, scientific data integrity, and database rectification and traceability.